様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ５ 月 ２９ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５０００２８

研究課題名（和文）文字列解析によるウェブソフトウェア開発支援

研究課題名（英文）String Analysis for the Development of Web Software

研究代表者
南出　靖彦（Minamide, Yasuhiko）

筑波大学・システム情報系・准教授

研究者番号：５０２５２５３１

交付決定額（研究期間全体）：（直接経費）　　3,900,000円

研究成果の概要（和文）：HTML5構文解析の信頼性の向上を目指した研究を行った．仕様に基づく網羅的なテストの自動生成によるHTML5構文解析器の仕様への適合性検査を実現した．形式化した仕様を条件付きプッシュダウンシステムへ変換し，プッシュダウンシステムに対し到達可能性解析を適用することで，テストの自動生成を行った．
ウェブソフトウェアに対するプログラム解析の精度を高めるために，正則表現マッチングの意味論の研究を行い，リストモナドに基づく意味を与えた．また，正規表現に対して，その正規表現によるマッチングの計算時間が，入力文字列の長さに対して線形であるかを判定する手法を木トランスデューサに理論に基づき構築した．

研究成果の概要（英文）：To improve the reliability of HTML5 parsers, we check the conformance of HTML5 parsers by applying automated test generation. We translate a formalized specification into a conditional pushdown system and apply the reachability analysis of pushdown systems to automatically generate test cases.
To improve the precision of program analysis for Web Software, we give the semantics of regular expression matching based on list monads. We also develop a method that checks whether the matching of a regular expression runs in linear time. The method is based on the theory of tree transducers.

研究分野：ソフトウェア検証

キーワード：ソフトウェア検証　プログラム解析　ウェブ　文脈自由文法　プッシュダウンシステム　HTML5

１．研究開始当初の背景

ウェブソフトウェアは，ウェブブラウザで解釈・実行される HTML 及び JavaScript，データベース上で実行される SQL 問合せを，サーバ側プログラムで動的に生成する技術によって非常に柔軟性なシステムとなっている．しかし，動的にスクリプトを生成することによって得られる柔軟性は，ウェブシステムの構成要素（サーバ，ブラウザ，データベース）間のインタラクションの不整合の原因となることがある．このようなに不整合よって生じる脆弱性（SQL インジェクション脆弱性，クロスサイトスクリプティング脆弱性）が，ウェブソフトウェアの信頼性を損なう重大な問題となっている．

これまでのウェブソフトウェアの開発においては，コードインスペクションやテストなどによって，脆弱性などの問題を見つけるアプローチがとられてきた．しかし，このような手法では，安全性を保証することは困難であり，新たなアプローチが求められている．また，ウェブソフトウェアの信頼性の向上においては，その基盤となるウェブブラウザ等の信頼性の向上も重要となる．例えば，ウェブブラウザにおける構文解析器の実装上の問題が想定外のクロスサイトスクリプティング脆弱性の原因となることも報告されている．

一方，本研究の研究代表者は，先行研究において，ウェブソフトウェアの検証の基盤として，プログラムの文字列出力を文脈自由文法を用いて，保守的に近似するプログラム解析（文字列解析）を開発した．このプログラム解析をサーバサイドプログラムに適用することで，生成されうるウェブページの近似を得ることができ，サーバサイドプログラムの脆弱性の検出や生成されるウェブページの妥当性検査が可能となっている．このようなソフトウェア検証の技術をウェブソフトウェア開発に統合し，ウェブソフトウェアの信頼性を高める開発手法・ツールを開発することが期待されている．

２．研究の目的

本研究は，ソフトウェア検証の技術をウェブソフトウェアの開発に適用し，ウェブソフトウェアの信頼性を高めることを目的とする．特に，出力付きオートマトン（トランスデューサ）や文脈自由言語などの形式言語理論に基づく検証技術のウェブソフトウェアへの適用を進める．

（１）ウェブソフトウェアにおいては，正規表現を用いた文字列操作がセキュリティに関連する検査などで重要な役割を果たしている．これまでの文字列解析では，正規表現を用いた文字列操作を正確に扱うことができていなかった．本研究では，まず，正規表現によるマッチングの振る舞いを明らかにする意味論を確立し，その意味論をウェブのためのプログラム解析の精度の向上の基礎とする．

（２）ウェブページの記述言語の最新仕様 HTML5 が広く利用され始めている．HTML5 では，構文解析アルゴリズムを詳細に規定することによって，これまで問題となってきたウェブブラウザにおける構文解析の非互換性の問題を解決することを目指している． この構文解析仕様はスタック機械の遷移として記述されているが，エラー処理等のため非常に複雑になっており，6000 行以上にもなる．また，記述が自然言語でなされているため，仕様が明確でない箇所も多く見られる．このような問題から，実際には，ウェブブラウザや構文解析ライブラリにおいて非互換性が生じている．本研究では，ソフトウェア検証の技術を HTML5 構文解析に適用し，仕様や実装の信頼性の向上を目指す．

（３）ウェブソフトウェアにおいて重要な役割を果たす正規表現マッチングについて，その計算量を検査する技術を開発する．正規表現マッチングは，最悪の場合，指数関数的な時間計算量を持つことが知られており，ReDoS(Regular expression Denial of Service)脆弱性と呼ばれる DoS 脆弱性の原因となる． 本研究では，形式言語理論に基づく ReDoS 脆弱性の検査手法を開発する．

３．研究の方法

ウェブソフトウェアにおける脆弱性など現実的な問題の分析と形式言語理論に基づく検証技術の研究を同時に進める．また，検証手法等の実験による実用性の評価には，本研究の代表者がこれまで開発してきた PHP 文字列解析器の枠組みを用いる．

本研究の基礎となる形式言語理論の道具としては，文字列や木構造上のトランスデューサが，まず，第一に上げられる．トランスデューサについては，広範かつ難解な研究成果が得られている．しかし，これまでの研究は，応用をほとんど考慮せず展開されており，検証への応用においては，結果の整理や新たなアルゴリズムの開発が必要となる．

また， Streaming String Transducer などの新たな計算モデルの検証への応用の可能性も探る．

４．研究成果

(1)ウェブソフトウェアに対するプログラム解析の精度をたかめるために，正則表現マッチングの意味論の研究を行った．正規表現マッチングを非決定的な構文解析器と考え，リストモナドを用いた操作的意味を与えた．また，正規表現マッチングにおける部分マッチの意味を，出力モナドを導入するモナド変換（monad transformer）を用いて自然に表現できることを示した．また，この意味論をプログラミング言語における正規表現の拡張

であるアトミックグループに拡張し，アトミックグループを含む正規表現が表す言語が正則であることを示した．

(2) 本研究では，HTML5 の構文解析プログラムの信頼性の向上を目指した研究を行った．特に，仕様に基づく網羅的なテストの自動生成によって，ウェブブラウザにおける実装の仕様への適合性（conformance）の向上を目指した．テストの自動生成は，形式化した仕様に対する到達可能性解析に基づいている．まず，仕様記述のための言語を導入し，構文解析仕様のサブセットを形式化した．次に，形式化した仕様を条件付きプッシュダウンシステムと呼ばれるプッシュダウンオートマトンの拡張へ変換した．条件付きプッシュダウンシステムはスタックの内部を正則言語で検査することができる．このような検査を HTML5 構文解析仕様は本質的に用いている．変換によって得られた条件付きプッシュダウンシステムに対し到達可能性解析を適用することで，テストの自動生成を行った．この手法により自動生成したウェブ文書により，ウェブブラウザや構文解析ライブラリを検査し，Safari，Firefox などのブラウザ及び主要な構文解析ライブラリにおける非互換性の発見に成功した．
また，本研究では，条件付きプッシュダウンシステムの到達可能性解析に関して，既存のアルゴリズムとは異なるアイデアに基づく，より効率的なアルゴリズムを開発した．さらに，このアルゴリズムを一般化し，重み付きプッシュダウンシステムの枠組みを拡張し，様々なプッシュダウンシステムの到達可能性解析がこの枠組みで実現できることを示した．
仕様自体の改善にも貢献した．この研究の過程で，構文解析仕様の最も複雑な部分が，HTML の構文解析として適切でない振る舞いを示すことが分かった．この問題点を HTML5 仕様の策定者に報告し，その結果，現在の仕様ではこの問題点が修正されている．

(3)プログラミング言語における正規表現の多くの実装は，バックトラックに基づいている．そのため，正規表現マッチングにかかる時間が文字列の長さに関して線形でないことがあり， 最悪の場合，指数関数時間となる．
本研究では，正規表現に対して，その正規表現によるマッチングの計算時間が，入力文字列の長さに対して線形であるかを判定する手法を提案する．検査対象の正規表現から，文字列を正規表現マッチングの計算過程を表す木に変換する先読み付きトップダウン木トランスデューサを構成する．この木トランスデューサで出力される木の大きさは，正規表現マッチングの計算時間に比例する．そのため，構成した木トランスデューサが入力の大きさに対して，定数倍の大きさに収まる木しか出力しないという性質を持つかを検査することで，計算量が文字列の長さに対して線形であることを判定することが出来る． この性質の検査には，Engelfriet と Maneth の方法を利用した．
この検査手法を OCaml によって実装し，既存の PHP プログラムで使用されている正規表現を対象に実験を行った．実験の結果，対象の正規表現 393 個中 47 個が非線形であると判定された．

(4)ウェブのためのプログラム解析の基盤として，新しい計算モデルである Streaming String Transducer に関する研究を行った．本研究では，関数的非決定性ストリーミング文字列トランスデューサの等価性判定を正規表現による文字列置換の等価性判定に応用した．ここで，関数的とは非決定性であっても出力が高々1 つであることを示す．関数的 SST の等価性判定問題の決定可能性は Alur らによって証明されている．しかしながら，判定方法は複雑であり直感的な理解が困難であった．そこで，本研究では関数的 SST の等価性判定アルゴリズムの簡略化を行った．正規表現による文字列置換はグループ変数にマッチした箇所を複数回出力することができることから有限状態トランスデューサでは模倣できない．そこで，正規表現による文字列置換を関数的 SST で模倣することにより，関数的 SST 上の検証問題に帰着させた．この手法により，正規表現による文字列置換の等価性判定器を実装し，オープンソースの PHP プログラム中に現れる実際の文字列置換に対して実験を行った．

５．主な発表論文等

〔雑誌論文〕（計８件）

① 加賀江 優幸，南出 靖彦，Streaming String Transducer の等価性判定と正規表現による文字列置換への応用，情報処理学会論文誌 プログラミング，査読有，採録決定

② 上里 友弥 ，南出 靖彦，スタック長の特徴付けによる言語の非 DCFL 性証明，情報処理学会論文誌 プログラミング，査読有，Vol.7，No.4，8-20，2014．
http://id.nii.ac.jp/1001/00102870/

③ Satoshi Sugiyama , Yasuhiko Minamide, Checking Time Linearity of Regular Expression Matching Based on Backtracking，情報処理学会論文誌 プログラミング，査読有,Vol.7，No.3，pp.1-11，2014．
http://id.nii.ac.jp/1001/00102152/

④ Yuya Uezato and Yasuhiko Minamide, Pushdown Systems with Stack

Manipulation, In Proc. the 11th International Symposium on Automated Technology for Verification and Analysis, 査読有,LNCS 8712, pp.412-426, 2013.
DOI: 10.1007/978-3-319-02444-8_29

⑤ Yasuhiko Minamide, Weighted Pushdown Systems with Indexed Weight Domains, In Proc.the 19th International Conference on Tools and Algorithms for the Construction and Analysis of Systems, 査読有,LNCS 7795, pp.230-244, 2013.
DOI: 10.1007/978-3-319-02444-8_29

⑥ 杉山 聡, 南出 靖彦, アトミックグループで拡張された正規表現のオートマトンへの変換, 情報処理学会論文誌 プログラミング, 査読有,Vol.6, No.1, pp.17-26, 2013.
http://id.nii.ac.jp/1001/00089431/

⑦ Yasuhiko Minamide, Shunsuke Mori, Reachability Analysis of the HTML5 Parser Specification and Its Application to Compatibility Testing, In Proc. the 18th International Symposium on Formal Methods, 査読有,LNCS 7436, pp.293-307, 2012
DOI: 10.1007/978-3-642-32759-9_26

⑧ Yuto Sakuma, Yasuhiko Minamide, Andrei Voronkov, Translating Regular Expression Matching into Transducers, Journal of Applied Logic, 査読有,10, pp. 32-51, 2012.
DOI: 10.1109/SYNASC.2010.50

〔学会発表〕(計７件)

① Yasuhiko Minamide, Complexity Analysis of Regular Expression Matching Base on Backtracking, Dagstuhl Seminar, Scripting Languages and Frameworks: Analysis and Verification, 2014 年 6 月 29 日～7 月 4 日, Schloss Dagstuhl, (ドイツ).

② Yasuhiko Minamide, HTML5 Parser Specification and Automated Test Generation, Dagstuhl Seminar, Scripting Languages and Frameworks: Analysis and Verification, 2014 年 6 月 29 日～7 月 4 日, Schloss Dagstuhl (ドイツ).

③ 上里 友弥 , 南出 靖彦, Conditional Transformable Pushdown System: スタックの変換と検査が可能なプッシュダウンシステム. 第 15 回プログラミングおよびプログラミング言語ワークショップ（PPL2015）, 2013 年 3 月 4 日～3 月 6 日. 道後プリンスホテル（愛媛県松山市）.

〔その他〕
本研究に関連する受賞

① 上里 友弥, 南出 靖彦.日本ソフトウェア科学会 プログラミング論研究会 第 15 回プログラミングおよびプログラミング言語ワークショップ 論文賞, 2013 年 3 月, Conditional Transformable Pushdown System: スタックの変換と検査が可能なプッシュダウンシステム

② 杉山聡, 情報処理学会 2014 年度山下記念研究賞,「バックトラックによる正規表現マッチングの時間計算量線形性判定」

③ 上里友弥, 情報処理学会 2014 年度コンピュータサイエンス領域奨励賞,「スタック長の特徴付けによる言語の非 DCFL 性証明」

④ 杉山聡, 情報処理学会 2013 年度コンピュータサイエンス領域奨励賞,「アトミックグループで拡張された正規表現のオートマトンへの変換」

６．研究組織
(1)研究代表者
南出 靖彦 (MINAMIDE, Yasuhiko)
筑波大学・システム情報系・准教授
研究者番号：50252531